University of Zurich[UZH]

Zurich Open Repository and Archive
University of Zurich
University Library
Strickhofstrasse 39
CH-8057 Zurich
www.zora.uzh.ch

Year: 2019

Lip Reading Deep Network Exploiting Multi-Modal Spiking Visual and Auditory Sensors

Li, Xiaoya ; Neil, Daniel ; Delbruck, Tobi ; Liu, Shih-Chii

Abstract: This work presents a lip reading deep neural network that fuses the asynchronous spiking outputs of two bio-inspired silicon multimodal sensors: the Dynamic Vision Sensor (DVS) and the Dynamic Audio Sensor (DAS). The fusion network is tested on the GRID visual-audio lipreading dataset. Classification is carried out using event-based features generated from the spikes of the DVS and DAS. Networks are trained separately on the two modalities and also jointly trained on both modalities. The jointly trained network when tested on DVS spike frames alone, showed a relative increase in accuracy of around 23% over that of the single DVS modality network.

DOI: https://doi.org/10.1109/iscas.2019.8702565

Posted at the Zurich Open Repository and Archive, University of Zurich
ZORA URL: https://doi.org/10.5167/uzh-184220
Conference or Workshop Item
Accepted Version

Originally published at:
Li, Xiaoya; Neil, Daniel; Delbruck, Tobi; Liu, Shih-Chii (2019). Lip Reading Deep Network Exploiting Multi-Modal Spiking Visual and Auditory Sensors. In: 2019 IEEE International Symposium on Circuits and Systems (ISCAS), Sapporo, Japan, 26 May 2019 - 29 May 2019, Institute of Electrical and Electronics Engineers.
DOI: https://doi.org/10.1109/iscas.2019.8702565

Accepted for publication at the 2019 IEEE International Symposium
on Circuits and Systems

Lip Reading Deep Network Exploiting Multi-modal Spiking Visual and Auditory Sensors

Xiaoya Li, Daniel Neil, Tobi Delbruck, and Shih-Chii Liu
Institute of Neuroinformatics, University of Zurich and ETH Zurich, Zurich, Switzerland
tobi, shih@ini.uzh.ch

***Abstract*—This work presents a lip reading deep neural network that fuses the asynchronous spiking outputs of two bio-inspired silicon multimodal sensors: the Dynamic Vision Sensor (DVS) and the Dynamic Audio Sensor (DAS). The fusion network is tested on the GRID visual-audio lipreading dataset. Classification is carried out using event-based features generated from the spikes of the DVS and DAS. Networks are trained separately on the two modalities and also jointly trained on both modalities. The jointly trained network when tested on DVS spike frames alone, showed a relative increase in accuracy of around 23% over that of the single DVS modality network.**

***Index Terms*—deep learning network, sensor fusion, spiking silicon sensors, spiking cochlea, Dynamic Vision Sensor,**

I. Introduction

Neuromorphic event-based sensors, in particular the visual and auditory sensors, have spurred the development of both event-based algorithms and spiking networks for real-time deployment. The most developed sensor is the Dynamic Vision Sensor (DVS) [1] leading to later variants such as the higher sensitivity DVS [2], DVS with spike encoding using an asynchronous delta sigma modulator [3], and retinas that produce both an event-driven output and intensity output such as the ATIS [4] and the DAVIS [5]. The most developed spiking silicon cochlea is the Dynamic Audio Sensor (DAS) [6]–[8].

These sensors have recently been combined together with both spiking and analog Deep Neural Networks (DNNs) for various machine learning tasks. Convolutional Neural Networks (CNNs) have been used to process the spikes or event-driven frames [9], [10]. Recurrent Neural Networks (RNNs) with the Long Short-Term Memory (LSTM) [11] and Gated-Recurrent Unit (GRU) [12], have been used successfully in speech recognition tasks [13]–[15]. These event-based systems show competitive performances in particular, low latencies and low power consumption, for real-world tasks.

The investigation of DNNs together with multimodal spiking sensors is still relatively rare. Previous studies of sensor fusion using DNNs on multimodal spiking sensors include a spiking Deep Belief network with the DVS and DAS [16], hardware equivalents for inference [17], and analog CNNs and RNNs using event-driven spike features [15], Another study used event cameras with audio input on voice recognition [18].

In speech recognition, the combination of audio and visual inputs gives better accuracies especially if one sensor type is noisy or less informative. In particular, deep network architectures are widely used for this task due to its ability in feature learning [19]. Between the two modalities, recognition with visual inputs is more difficult, e.g. in lipreading. State-of-art approaches using deep learning on a lip reading task show that the trained network accuracy can exceed human professionals [20]–[22].

This work presents a lip-reading DNN tested on a word level speech recognition problem from event-based data. The dataset used is the GRID corpus [23], an audio-visual benchmark dataset. The fusion network includes two modalities and takes an audio and the corresponding video as input. Four different inputs were tested: analog audio frames, video frames; and frames from spiking event recordings of the original GRID dataset. In addition, single modality networks with only the video or audio branch are also tested. The remainder of the paper is organized as follows: The details of the network architectures and the feature extraction methods used on the different input representations are described in Section II; the experiments on the single and dual modality networks are presented in Section III and finally, the results are discussed in Section IV.

II. Methods

The different sensors used in the recordings, the deep network architectures, the recording details of the GRID spiking dataset and the preprocessing of the dataset are described next.

A. Spiking Multi-modal Sensors

a) Dynamic Vision Sensor: The DAVIS 240C camera [5] is used for the visual spike recordings. This sensor provides both DVS event and APS frame outputs. The DVS output includes both ON and OFF events, corresponding to the polarity of the contrast change at each pixel and their timestamps.

b) Dynamic Audio Sensor: The Dynamic Audio Sensor used in this workhas 64 channels, each corresponding to a central frequency in the range of 20 Hz to 5 kHz. The system models the bandpass filtering of the basilar membrane of the biological cochlea. It then generates spike events from delta changes in the rectified channel output using an asynchronous delta modulation scheme. It encodes the audio inputs by ON and OFF events but only the ON events are used in this work.

This work was partially funded by the Swiss National Science Foundation Grant HEAR-EAR, 200021_172553.

B. Dataset Preprocessing

The GRID audiovisual sentence corpus [23] used in this work, consists of audio and video (facial) recordings of 1000 sentences spoken by each of 34 talkers (18 male, 16 female). Each sentence consist of a six word sequence of the form of command + color + preposition + letter + digit + adverb, for example, "put red at G 9 now". There are in total 51 different words in the GRID corpus. The letter w was excluded in the corpus since it is multisyllabic.

In order to preserve the details in the subjects lip movements, the facial area from each video frame is extracted through the OpenCV face detector [24] with an example shown in Fig. 1.

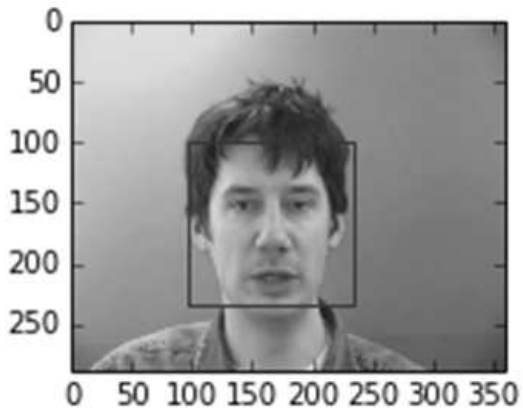

Fig. 1: Cropped region of an example video frame from the GRID corpus.

C. Spiking Sensor Recording Setup

The DAVIS camera with a 4.5mm lens is placed 20cm away from the screen of an AOC i2369Vm 1920 x 1080p LED monitor. The audio files are played from a sound card to the DAS.

The frame output of the DAVIS camera is used to adjust both the camera position and lens focus prior to recording. The frame output is turned off during recording because of disk storage limitations due to the large uncompressed files of the captured frames. The cropped frames from the GRID video are enlarged so that the figure fills up the screen. This was done because the DAVIS resolution is only 340$\times$ 280 and the finer lip movements needs to be enlarged for the camera to capture more details.

In total, audio and video spikes from 21k sentences from speakers 1-10, 16-20 and 22-30 are used in the network experiments. The remaining videos are not used due to the following reasons: 1) Software and/or hardware failure during re-recording; 2) the face extractor failed to locate the face in the video; and 3) videos of speaker 21 are not available.

a) Audio frame features: Mel Frequency Cepstral Coefficient (MFCC) features are extracted from each frame. The 39-dimensional vectors include both 1st and 2nd order derivatives. The window size is set to 60ms and the frame shift to 40ms. The 40ms frame shift is chosen so that the shift is consistent with the frame rate (25 fps) of the video.

b) Video frame features: The extracted 180 $\times$ 180 frames from the OpenCV face detector as applied to the original video frames of the GRID dataset, are then downsampled to 48$\times$48 using bicubic interpolation. The original 25fps frame rate is preserved.

c) Audio spike input: Spike count features are generated from each channel of the DAS by using only the ON events and a 40ms window and zero frame shift [25]. A vector of length 64 (from the 64 channels) is generated at each time step t. The spike count vectors at all time steps are then concatenated to form the input sample.

d) Video spike features: The event stream from the DAVIS 240C recordings is binned using a 40ms window. The value of each pixel in the binned frame is set by the net number of ON events minus OFF events counts. This method has been shown to be beneficial for normalization [9]. The binned frames were then resized to 48$\times$48 resolution using bicubic interpolation. Figure 2 shows binned spike events frames from a video sample after resizing (a) and normalization (b).

D. Network Architecture

The sensor fusion network consists of an audio feature extraction network, a video feature extraction network and post merge classification layers. The audio features are extracted by a single layer 150-GRU RNN layer [12]. Because the feature vector sizes of MFCC and DAS spikegram are different, the input weight matrix dimension of the GRU layer is 39 x 150 for the MFCC inputs and 64 x 150 for the audio spikegram inputs. The video features are computed by a 3-layered CNN and a single 80-unit GRU layer. The CNN layers consist of a single 5 x 5 x 8 convolutional layer followed by a 2 x 2 max pooling layer, stacked three times. The outputs of the audio RNN and video RNN are concatenated along the time axis to form the input to the classification layers. In each training or testing batch, the audio and video inputs are set to the same length for all samples of the batch by padding either the MFCC frames or the video frames with zeros. The classification layers are composed of one 240-unit GRU layer, one 250-unit fully connected (FC) layer and one 250 x 51 (FC) layer. The output of the network is a 51-dimensional multi-hot vector that corresponds to the 51 available words. The cost function computes the total mismatch between the network output vector and target vector divided by the number of words in the target. The input to the classification layers can be the concatenated features from the two single modality networks in the case of the fusion network, or the features from a single modality network. In the latter case, the sensor fusion network becomes a single modality network.

E. Training

The dataset is divided randomly into a training set containing 90% of the sentences; and a testing set containing the remaining 10%. The division of the training and testing sets is speaker independent. The training is done using the Adam optimizer with a learning rate of 5×10^{-4} for all experiments.

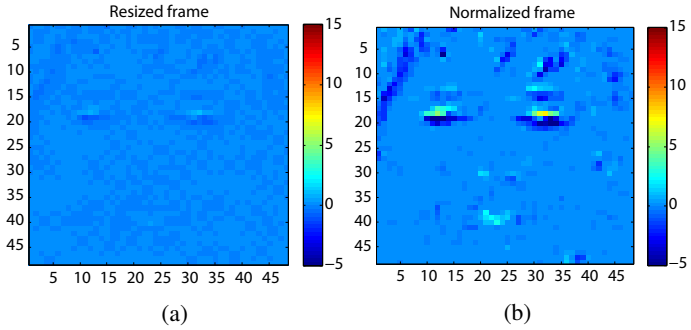

Fig. 2: (a) Resized frame. (b) Normalized frame.

Different amounts of blankout are applied on both the audio and video input to the merge network to deal with the fact that the audio input is more informative.

F. Evaluation Metric

The network receives whole sentences as input, but the classification and evaluation is done at the word level. The sentences are translated into bag-of-words (BoW) representation using a vocabulary size of 51. Due to the strict sentence structure used in the GRID corpus, each word can only appear once in a sentence. Therefore instead of counting the number of word occurrences in the BoW representation, we use a binary 51-dimensional vector to represent a sentence. The position corresponding to each word is set to be 1 if the word appears in the sentence, and 0 otherwise. The outputs of a network are analog vectors of length 51, which are then converted to the same BoW vector format by thresholding with a value of 0.5. The word accuracy in percent, WA, is computed by calculating the distance between the target sentence vector and network prediction following:

$$(1 - N_{wg}/N_t) \ * 100\% \tag{1}$$

where N_{wg} is the number of wrong guesses and N_t is the number of words in the target. N_{wg} is the sum of the number of wrong words, missed words and extra words in the prediction. We assume that there is no repeated word in a prediction. Because the maximum number of wrong guesses is 51, the range of possible classification accuracy is [-7.5,1] given that $N_t = 6$.

G. Correlation between Event-based Sensor Outputs

In general, the DVS and DAS events from the same sample are not of the same length and are not aligned even though the same start signal for recording is given to both sensors simultaneously. Spikes are generated from the DAS sensor only starts when speech begins. The recorded DVS events are noisy also contain a large portion of spikes that are not related to lip movements, for example, events generated by the speakers movements. The recorded events also contain background spikes due to the dark current of the pixel photodiode.

To filter out the noisy spikes and to align both modality inputs, we look for DVS events that are highly correlated

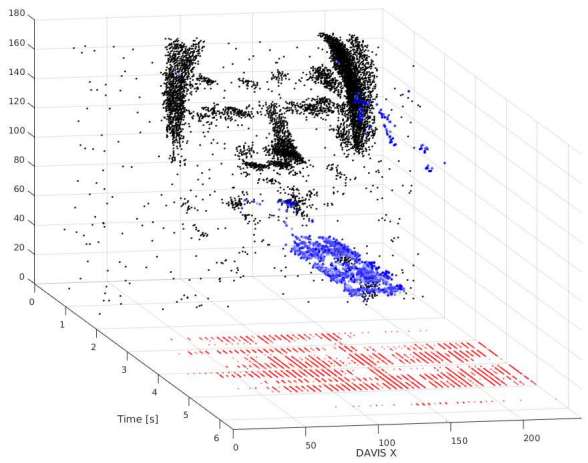

Fig. 3: Figure shows raw DVS spikes in black, the correlated DVS spikes in blue, that are correlated with the audio DAS spikes in red.

in time with DAS events in each sample. These DVS events were selected by running the correlation function on the DVS spike frames with the DAS spike frames from the duration of the sample. Events from spike frames with correlation values lower than 0.2 are removed. The original and DVS spikes from the filtered frames are shown in Fig. 3. Due to the processing time of the correlation and filtering operation, only 5000 samples were used for the experiments.

III. Experimental Results

The results for the single modality and sensor fusion networks are presented next.

A. Single Modality Network

The classification accuracy of a network receiving a single input modality is presented in Table I. The experiments are carried out on the same dataset of 19k training samples and 2k validation samples. The accuracies for the single modality networks are used as a comparison with the performance of the sensor fusion network. The RNN using MFCC features already gives an accuracy of around 98.41%. The audio inputs yield better performance than the corresponding video inputs (accuracy around 84.27%), which is expected as the audio is more informative than lip movement for this task. The test accuracy of the video frame network is similar to the performances from other networks which are trained on the GRID dataset, e.g. [21].

TABLE I: Measured accuracy on single modality network.

Network Input	Word Accuracy
RNN for DAS spike frames	83.83%
CNN+RNN for DVS spike frames	38.26%

In terms of training speed measured through epochs, the training of the network using audio MFCC features is faster

than the network using video inputs most likely because the accuracy from the audio input is much higher than the video input. This higher accuracy is also true of DAS spike inputs vs DVS spike inputs.

B. Performance of Sensor Fusion Network

Two types of fusion inputs are tested, DAS spike frames and video frames; and DAS and DVS spike frames. The results in Table II are obtained using a blankout policy on both inputs so that the network does not depend only on the more informative audio modality. The network trained jointly on both DAS and DVS spike frames, produced a word accuracy (WA) that lie between the accuracies from the two single modality networks. One reason is that only a subset of the dataset ends up being used for the network training due to the blank out policy in the joint training. More specifically, only 60% of the audio samples and 90% of the video samples are presented to the network during the joint training.

The accuracy results from the use of correlated and non-correlated DAS and DVS spike frames in the jointly trained network show that the filtered correlated inputs yield a better overall accuracy, even though the number of samples used for the correlated dataset is smaller than the number of samples in the uncorrelated dataset. The network is also tested on single modality inputs. This testing is done by setting all samples from the other modality to zeros. The audio word accuracy decreased significantly probably because of the higher reduction of audio samples over the video samples from the blankout policy. The jointly trained network when tested only on the DVS spike frames show an accuracy of 61.64% which is a big increase from the accuracy (38.26%) of the single video modality network.

TABLE II: Measured accuracy from sensor fusion network. * indicates results from the filtered correlated DVS and DAS spikes.

Fusion Network Input	Word Accuracy
DAS and DVS spike frames	72.67%
DAS and DVS spike frames*	86.66%

TABLE III: Measured accuracy of fusion network tested on either DVS or DAS spike frames alone. WA - word accuracy.

Trained Fusion Network Input	Audio WA	Video WA
DAS and DVS spike frames	63.04%	61.64%

IV. Conclusion

This work presents a study of a multi-modal fusion deep network on event-based viusal-audio spiking sensors using a lip reading dataset. It extends past previous audio-visual spiking sensor fusion studies where the audio input is either composed of simple combination tones or separate datasets are used for the different modalities [15]. The network in this work was trained using an audio-visual lip reading dataset. Both single modality networks and fused networks were trained on the audio and visual frames.

The single modality networks that were trained on separately on the DAS and DVS spike frames achieved lower accuracy than single modality networks that were trained on the audio MFCC features and the video frames. The input processing methods for the spiking sensors can still be improved, for example, the DVS spike frames are not as sharp as the original video frames. Different feature extraction methods can also be employed such as constant-event features. The fusion network however produces an increase in accuracy of 23% compared to a network trained only on DVS spike frames demonstrating that the DAS spike frames helped to improve the performance of the single video modality network. Future work will extend to direct recordings with both sensors on a similar task. Such recordings will allow one to study the use of the finer resolution temporal information that could be carried by both spiking sensor modalities.

Acknowledgment

The authors acknowledge P. Park for help with the combined sensor recordings and Y. Hu for help with the code.

References

[1] P. Lichtsteiner, C. Posch, and T. Delbruck, "A 128 × 128 120 dB 15 μs latency asynchronous temporal contrast vision sensor," *IEEE Journal of Solid-State Circuits*, vol. 43, no. 2, pp. 566–576, Feb 2008.

[2] T. Serrano-Gotarredona and B. Linares-Barranco, "A 128 × 128 1.5% contrast sensitivity 0.9% FPN 3 μs latency 4 mW asynchronous frame-free Dynamic Vision Sensor using transimpedance preamplifiers," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 3, pp. 827–838, March 2013.

[3] M. Yang, S. C. Liu, and T. Delbruck, "A Dynamic Vision Sensor with 1% temporal contrast sensitivity and in-pixel asynchronous delta modulator for event encoding," *IEEE Journal of Solid-State Circuits*, vol. 50, no. 9, pp. 2149–2160, Sept 2015.

[4] C. Posch, D. Matolin, and R. Wohlgenannt, "A QVGA 143 dB dynamic range frame-free PWM image sensor with lossless pixel-level video compression and time-domain CDS," *IEEE Journal of Solid-State Circuits*, vol. 46, no. 1, pp. 259–275, Jan 2011.

[5] C. Brandli, R. Berner, M. Yang, S.-C. Liu, and T. Delbruck, "A 240× 180 130 db 3 μs latency global shutter spatiotemporal vision sensor," *IEEE Journal of Solid-State Circuits*, vol. 49, no. 10, pp. 2333–2341, 2014.

[6] M. Yang, C. Chien, T. Delbruck, and S. Liu, "A 0.5v 55 μW 64 × 2 channel binaural silicon cochlea for event-driven stereo-audio sensing," *IEEE Journal of Solid-State Circuits*, vol. 51, no. 11, pp. 2554–2569, Nov 2016.

[7] S. Liu, A. van Schaik, B. A. Minch, and T. Delbruck, "Asynchronous binaural spatial audition sensor with 2 × 64 × 4 channel output," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 8, no. 4, pp. 453–464, Aug 2014.

[8] V. Chan, S. C. Liu, and A. van Schaik, "AER EAR: A matched silicon cochlea pair with address event representation interface," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 54, no. 1, pp. 48–59, Jan 2007.

[9] D. P. Moeys, F. Corradi, E. Kerr, P. Vance, G. Das, D. Neil, D. Kerr, and T. Delbrück, "Steering a predator robot using a mixed frame/event-driven convolutional neural network," in *2016 Second International Conference on Event-based Control, Communication, and Signal Processing (EBCCSP)*, June 2016, pp. 1–8.

[10] J. A. Pérez-Carrasco, B. Zhao, C. Serrano, B. Acha, T. Serrano-Gotarredona, S. Chen, and B. Linares-Barranco, "Mapping from frame-driven to frame-free event-driven vision systems by low-rate rate coding and coincidence processing–Application to feedforward ConvNets," *IEEE Trans on Pattern Analysis and Machine Intelligence*, vol. 35, no. 11, pp. 2706–2719, 2013.

[11] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997. [Online]. Available: http://dx.doi.org/10.1162/neco.1997.9.8.1735

[12] K. Cho, B. van Merriënboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN Encoder–Decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Oct. 2014, pp. 1724–1734. [Online]. Available: http://www.aclweb.org/anthology/D14-1179

[13] A. Amir, B. Taba, D. J. Berg, T. Melano, J. L. McKinstry, C. Di Nolfo, T. K. Nayak, A. Andreopoulos, G. Garreau, M. Mendoza *et al.*, "A low power, fully event-based gesture recognition system." in *CVPR*, 2017, pp. 7388–7397.

[14] W. Tsai, D. R. Barch, A. S. Cassidy, M. V. DeBole, A. Andreopoulos, B. L. Jackson, M. D. Flickner, J. V. Arthur, D. S. Modha, J. Sampson, and V. Narayanan, "Always-on speech recognition using truenorth, a reconfigurable, neurosynaptic processor," *IEEE Transactions on Computers*, vol. 66, no. 6, pp. 996–1007, June 2017.

[15] D. Neil and S. Liu, "Effective sensor fusion with event-based sensors and deep network architectures," in *2016 IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2016, pp. 2282–2285.

[16] P. O'Connor, D. Neil, S.-C. Liu, T. Delbruck, and M. Pfeiffer, "Real-time classification and sensor fusion with a spiking Deep Belief Network," *Frontiers in Neuroscience*, vol. 7, 2013.

[17] I. Kiselev, D. Neil, and S.-C. Liu, "Event-driven deep neural network hardware system for sensor fusion," in *2016 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2016, pp. 2495–2498.

[18] A. Savran, R. Tavarone, B. Higy, L. Badino, and C. Bartolozzi, "Energy and computation efficient audio-visual voice activity detection driven by event-cameras," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 2018, pp. 333–340.

[19] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 689–696.

[20] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, "Audio-visual speech recognition using deep learning," *Applied Intelligence*, vol. 42, no. 4, pp. 722–737, 2015.

[21] M. Wand, J. Koutník, and J. Schmidhuber, "Lipreading with long short-term memory," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 6115–6119.

[22] J. S. Chung, A. W. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild." in *CVPR*, 2017, pp. 3444–3453.

[23] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[24] G. Bradski and A. Kaehler, "OpenCV," *Dr. Dobbs journal of software tools*, vol. 3, 2000.

[25] J. Anumula, D. Neil, T. Delbruck, and S.-C. Liu, "Feature representations for neuromorphic audio spike streams," *Frontiers of Neuroscience*, 2018.